Use of whole genome sequencing to estimate the mutation rate of *Mycobacterium tuberculosis* during latent infection.

Christopher B. Ford[1,11], Philana Ling Lin[2,11], Michael Chase[1], Rupal R. Shah[1], Oleg Iartchouk[3], James Galagan[4,5,6], Nilofar Mohaideen[7], Thomas R. Ioerger[7], James C. Sacchettini[7], Marc Lipsitch[1,9], JoAnne L. Flynn[10], Sarah M. Fortune[1]

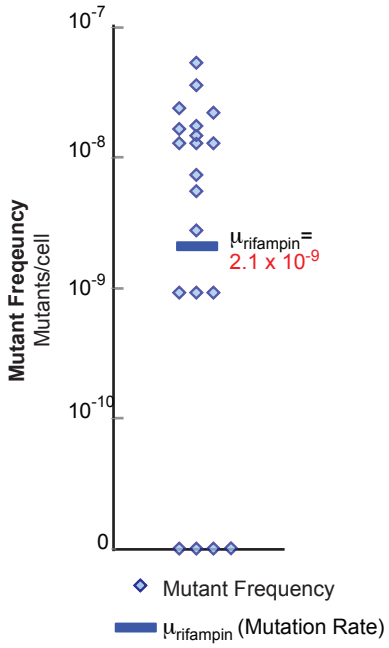
Supplementary Information

Supplementary Figure 1. The per base mutation rate of Mtb *in vitro*.

a

CFU	# of plates
0	4
1	3
3	1
6	1
8	1
14	3
16	1
18	1
19	1
24	1
26	1
39	1
58	1
Number of Cultures	20
MSS m value	2.27
Cell Count (N_t)	1.08×10^9
$\mu_{rifampin}$	2.1×10^{-9}
$\mu_{rifampicin}$ 95% CI	$1.01\text{-}3.30 \times 10^{-9}$

b



c

H37Rv Coordinate	Mutation	Amino Acid Change
761101	C>A	Q513E
761128	C>T	S522L
761128	C>G	S522W
761139	C>G	H526D
761139	C>T	H526Y
761140	A>G	H526R
761140	A>C	H526P
761149	G>T	R529L
761155	C>T	S531L
761155	C>G	S531W
Rifampin target size (bp)		10
$\mu_{in\ vitro}$		2.1×10^{-10}
$\mu_{in\ vitro}$ 95% CI		$1.01\text{-}3.30 \times 10^{-10}$

Supplementary Figure 1. The per base mutation rate of Mtb *in vitro*. (**a**,**b**) Luria and Delbrück fluctuation analysis was used to determine the rate of resistance to rifampicin. 20 independent cultures containing 1.08×10^9 cells each were plated and the resistance frequency was determined for each. The rate of resistance, $\mu_{rifampicin}$, was determined using the MSS method to calculate $m_{rifampicin}$ (m=2.27), the representative number of mutations per culture. These data are representative of four biologically independent experiments. In (**b**), diamonds represent individual cultures and their mutant frequency. The bar represents the mutation rate calculated from the frequency of mutants in each culture. (**c**) The number of mutations conferring rifampicin resistance in our assay was determined using Sanger sequencing. Sequencing rifampicin resistant isolates from 96 independent cultures identified ten unique mutations. The amino acid changes represent the standard codon annotation used in *E. coli*. The per base mutation rate, $\mu_{in\ vitro}$, was determined by dividing $\mu_{rifampicin}$ by the target size (10).

Supplementary Table 1. Coverage and read depth for each sequenced isolate.

Animal & Isolate	Run Identifier[a]	Read Length (bp)[b]	Average Read Depth	Percent Coverage[c]	Accession Number[d]
A – 1	15304-1B	75	278	96	SRR023458
A – 2	15304_2A	36	15	61	SRR122234
A – 3	15304-3A	75	25	86	SRR023468
A – 4	15304_4B	36	28	93	SRR122213
B – 1	7404-1	75	141	95	SRR023459
C – 1	7904-1	75	284	96	SRR023460
C – 2	7904-2	75	132	94	SRR023469
C – 3	7904-3	36	36	94	SRR122212
C – 4	7904-5	51	62	99	SRR122212
D – 1	11208-E4	51	15	97	SRR121590
D – 2	11208-E5	51	47	99	SRR121585
D – 3	11208-E7	51	43	99	SRR121586
D – 4	11208-J3	51	114	99	SRR121587
D – 5	11208-J6	51	76	99	SRR121588
D – 6	11208-J7	51	18	97	SRR121589
E – 1	7604-2	75	155	95	SRR023470
E – 2	7604-4	51	63	99	SRR023429
F – 1	8104-1C	75	148	89	SRR023462
F – 2	8104-2A	75	46	66	SRR023491
F – 3	8104_3B	36	22	91	SRR122235
G – 1	6404-1A	75	86	94	SRR023456
G – 2	6404-1B	75	183	95	SRR023461
G – 3	6404-3B	75	163	94	SRR023472
H – 1	11105-1	51	62	99	SRR121584
H – 2	11105-2	75	227	95	SRR023465
H – 3	11105-3	75	208	96	SRR023433
I – 1	10403-1	75	228	95	SRR023471
I – 2	10403-4	75	121	90	SRR023474
I – 3	10403-7	75	224	94	SRR023473
I – 4	10403-8	75	205	95	SRR023463
I – 5	10403-9	51	80	98	SRR121583
I – 6	10403-10	75	189	96	SRR023464
I – 7	10403-11	75	207	96	SRR023467
Inoculum: JF-Erdman	Library 1, Run 1	36	38	94	SRR121592
Inoculum: JF-Erdman	Library 1, Run 2	36	38	94	SRR121593
Inoculum: JF-Erdman	Library 2, Run 1	36	38	94	SRR121594
Inoculum: JF-Erdman	Library 2, Run 2	36	38	94	SRR121595

[a]Run Identifiers are based on the animal and strain identifiers used by Lin et al. JF-Erdman was run 4 times, twice from two separate libraries. Coverage values for JF-Erdman are based on pooled runs. [b]75bp reads were produced as paired-end reads by the Broad Institute of MIT and Harvard. Reads were subsequently trimmed to 48bp and pairing data was not used in assembly. 51 bp reads were produced by the Sacchettini lab at Texas A&M and were analyzed as paired end reads. 36bp reads were produced by Partners Healthcare Center for Personalized Genetic Medicine, Harvard Medical School. [c]Percent coverage values (defined as the fraction of sites covered in a genome) vary based on the method of assembly used. For 51bp reads, repetitive sequences were mapped using paired-end

data to disambiguate the location of reads that map to multiple locations in the genome, allowing for a higher percent coverage value. A read depth of five was required to call a SNP. For 75 and 36 bp reads, repetitive sequences that could map to multiple locations were discarded resulting in a lower percent coverage for these isolates and a read depth of ten was required to call a SNP. [d]Accession numbers refer to NCBI-SRA accession numbers where short reads have been posted.

Supplementary Table 2. Primers used in PCR/Sequencing of validated SNPs and in sequencing *rpoB*.

H37Rv Coordinate or Gene	SNP	Forward Primer	Reverse Primer
635497	C>T	GACGTCGTCACTACCAGGG	CTCGGTGACTTCGACCAGAT
690264	G>C	GCGATGGTGGTCAATCTGCTGCC	TGCTACCTCCCGTCCCGTCAG
693453	C>T	TGGTGGTCCTTGGTTGGTAT	TAGTCGTGGTGATCGTCTGC
766229	G>A	GACCCGTACATCGAAACCTC	AGACCACCGGTGATGTCCTC
975906	T>C	GCAAGTACATCCGAGAACCC	TTCCAATACTGCCGGAAGAC
1256717	G>A	GCGATCAGCTATCTCGGTG	GTAGATGTCGGATTGGGTCG
1854208	G>T	ACCCATACAACGGCAAGTGT	CGAGCGCTCTCATACAGACA
1861203	G>A	GAGGTTTCTCCCACCCTTACCGAC	CTCGGGACACGTTGCGCAC
2448250	G>A	CCCTCTAGGCTTGACGACAG	CGAGGTCTCGTAGGTCGGTA
3655598	G>T	GTGTCGACAAGCTGCATCAC	ACGATGGTGATGGCGTAGAT
4346906	C>A	GTACATGTTGATGATGCCGC	ACATATGACTGACCGGCTCC
rpoB	-	TCGGCGAGCTGATCCAAAACCA	ACGTCCATGTAGTCCACCTCAGA

The following polymorphisms were identified independently multiple times by WGS and were not subjected to PCR resequencing: 682043, 2350697, 4183984.